Preface: Computational Technologies for Drug Discovery

Pharmaceutical companies and technology companies have long enjoyed a close relationship, as the former constantly seek innovative new medicines and innovative new approaches to finding them. In this current era when data science and AI are increasingly touted for their potential to enhance and transform industries, drug discovery researchers should be recognized as early adopters of these approaches and savvy consumers and developers of new methods. Indeed, while the commercial output of a pharmaceutical company may be dozens of marketed drugs, a key internal output is the thousands or millions of data points that are generated as each new target is pursued in a cycle where relevant data is found and integrated, new data is generated by screening existing molecules, hypotheses are developed, new molecules are designed and synthesized, and the data analysis cycle starts anew. Such data represents a proprietary asset for the company a can be leveraged to increase innovation and efficiency.

During my 20 years in the pharmaceutical industry, I experienced firsthand the technical diversity and depth of a pharmaceutical company's "Computational Chemistry" group. As early as 1988, team members possessed a wide range of expertise—database management, statistics, organic chemistry, chemical synthesis prediction, small molecule crystallography, quantum chemistry, protein molecular dynamics, enzymology, and more. In addition, at that time, like now, the industry needed high-performance computing resources such as the IBM 3090 series to support the compute-intensive calculations.

This special issue of the *IBM Journal of Research and Development* focuses broadly on Computational Technologies for Drug Discovery and includes reflections from both IBM and non-IBM authors on the current and future state of the art for a number of key technologies (the "what") as well as descriptions of how, when, and why they are used.

The first paper, written by senior pharmaceutical industry veteran Kelvin Cooper, summarizes insights into the failure of drug pipeline candidates, a topic upon which pharmaceutical companies have become increasingly reflective over the past two decades and for which they and they alone have the data to evaluate trends, correlations, and causes. The three key reasons for failure were: poor physio-chemical properties of the drug molecules, insufficient efficacy of the chosen target, and inconsistent organizational strategy leading to changing priorities and sometimes cultural turmoil and distraction.

In the second paper, IBM researcher Edward Pyzer-Knapp describes Bayesian optimization for accelerated drug discovery, an approach that balances the strategies of exploiting known information and exploring new space when choosing which experiment to conduct next, for example when searching for new leads in chemical space. Pharmaceutical companies have long looked to their internal proprietary compound collections for new leads, but as collection sizes have expanded into the millions, an efficient way to search has become important to maximize efficiency and minimize unnecessary depletion of the physical compound collection.

In the third paper, David Koes describes the Pharmit online resource for interactive high-throughput drug discovery searches based on pharmacophore or 3-D shape models. The ever increasing availability of such online and open-source applications, as well as life science knowledge bases in the public domain, fosters the development of new methods and enables innovation for researchers who might otherwise lack sufficient data, software, or compute resources.

The fourth paper by Bilge Acun et al. presents algorithmic improvement and performance optimization for the NAMD molecular dynamics code as run on current supercomputers, namely the Oak Ridge National Laboratory Summit and Lawrence Livermore National Laboratory Sierra machines. These computers are IBM Newell platform (with IBM POWER9 processors and NVIDIA Volta V100 GPUs) and held the number one and number three spots in the June 2018 world ranking of top 500 supercomputers. Molecular dynamics simulations provide the conformational and configurational sampling required to calculate accurate free energies of binding to rank candidate small molecules for synthesis, understand structure activity relationships (SARs), and capture relevant conformational changes. Increases in computer speed support the investigation of ever larger biological systems, as well as the more accurate modeling of systems that are currently in scope.

In the fifth paper, Teodoro Laino et al. describe the implementation of an efficient algorithm to support quantum chemical protein molecular dynamics simulations for drug discovery as an alternative to the now standard classical approach. The large-scale parallel sparse matrix–matrix multiplication approach was applied to the semiempirical NDDO Hamiltonian, which was then easily parameterized to support calculation of inter- and intramolecular interactions in proteins. Recent decades have seen many major advances in sampling supported by faster computers and new algorithms, but less dramatic progress has been seen with the energy representations due to the complexity of representing these condensed phase systems.

The sixth paper, by Yudong Cao et al., focuses on the exciting new technology of quantum computing and suggests opportunities to impact approaches of relevance to drug discovery, such as machine learning and energy-based simulations. This novel hardware architecture invites

Digital Object Identifier: 10.1147/JRD.2019.2893982

researchers to consider new ways to formulate old use cases leveraging algorithms that most efficiently exploit the new compute capabilities.

In the seventh paper, Jaehee Shim et al. present a new approach to identifying beneficial drug–drug combinations from real-world evidence, specifically by looking for prediction patterns through the lens of drug classes. The authors demonstrate their approach on 78,345 drug–drug combination predictions made from the FDA Adverse Event Reporting System (FAERS). This is one of many recent examples of real-world evidence (RWE) that sources such as FAERS, electronic health records, and insurance claims being mined to identify relevant trends that are often not evident from preclinical or clinical studies due to the relative simplicity of those studies (e.g., single drug, limited co-morbidity, etc.) or small sample size.

In the final paper of this issue, Richard Martin et al. detail the information extraction process used within the IBM Watson for Drug Discovery cognitive computing platform, which searches large-scale unstructured data such as literature abstracts and full text content, as well as other unstructured and structured content sources, using both model- and rule-based techniques. The entities and relationships that are extracted support knowledge discovery and predictive analytics use cases in drug discovery, including target identification and validation. Although the days are past when drug discovery researchers could routinely look to individual papers in the scientific literature to provide novel validated targets to pursue, the aggregate analysis of this big data source can lead to otherwise hidden insights.

In conclusion, this issue contains a stimulating mix of internal and external contributions covering a range of topics relevant to drug discovery researchers. New data sources, new algorithms, new hardware, and new collaboration models all offer potential for the more efficient discovery of innovative new medicines. We hope you find the issue to be informative, enjoyable, and provocative.

Wendy Cornell
Global Strategy Lead and Manager, Drug Discovery Technologies
IBM Healthcare and Life Sciences Research
Guest Editor